MOTIF PREDICTION
Motifs

- Small, Variable, Repetitive Sequences
- Promoters for Transcription Factors
- Also:
  - Splice signals, Kinase Targets, Immune targets: MHC
- Predict Motif Instances
- Learn Motif Structure
A Promoter Model

The same motif model in all promoters
Probabilistic Model

Count frequencies
Add pseudocounts

Position Frequency Matrix (PFM)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEM13</td>
<td>.0</td>
<td>.0</td>
<td>.2</td>
<td>.9</td>
</tr>
<tr>
<td>HEM13</td>
<td>.5</td>
<td>.7</td>
<td>.5</td>
<td>.1</td>
</tr>
<tr>
<td>ANB1</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
</tr>
<tr>
<td>ANB1</td>
<td>.5</td>
<td>.3</td>
<td>.3</td>
<td>.0</td>
</tr>
<tr>
<td>ANB1</td>
<td>.5</td>
<td>.3</td>
<td>.0</td>
<td>.1</td>
</tr>
<tr>
<td>ROX1</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
</tr>
</tbody>
</table>
Parameterizing the Motif Model

Given multiple sequences and motif locations but *no motif model*

- AATGCG
- ATATGG
- ATATCG
- GATGCA

1. Count Frequencies
2. Add pseudocounts
Expectation Maximization Technique

- Hidden Data, stochastic model
- Start with PFM guess
- Estimate Parameters, which depend on hidden data
  - Marginalize observables over hidden data
  - Motif positions
- Update model – New PFM
  - Find parameters that maximize likelihood of the model
  - Likelihood of sequence with Motif positions, PFM
EM: Motif Discovery

- Don’t know starting positions of motif
- Don’t know PFM
- Randomly start

Form the Profile, P from s

Find Probability of Sequence
- Motif at starting point i=1…n-k+1
- Choose Best Start for all t
- Repeat with new starting points

\[ S = s_1, s_2, \ldots, s_t \]
EM for Motif Discovery

1. Start with random motif model
2. **E Step**: estimate probability of motif positions for each sequence
3. **M Step**: use estimate to update motif model
4. Iterate (to convergence)
The M-Step Calculating the Motif Matrix (Count)

- $M_{ck}$ is the probability of character $c$ at position $k$
- With specific motif positions, we can estimate $M_{ck}$:

Counts of $c$ at pos $k$
In each motif position

$$M_{c,k} = \frac{n_{c,k} + d_{c,k}}{\sum_b n_{b,k} + d_{b,k}}$$

$$n_{c,k} = \sum_{\text{sequences } S_i} \sum_{j|S_i=c} Z_{ij}$$
MEME

- Expectation-Maximization for motif finding
  - Proteins, DNA
- Starts at multiple points (Motif positions)
- Choose best score
- Because
  - Search space of starting points is large
  - Rarely close to optimal
- All Motif positions in all sequences updated at each iteration
$P(\text{Seq} \mid \text{Model})$ Landscape

EM searches for parameters to increase $P(\text{seqs} \mid \text{parameters})$

Useful to think of $P(\text{seqs} \mid \text{parameters})$ as a function of parameters.

EM starts at an initial set of parameters $\bigcirc$.

And then “climbs uphill” until it reaches a local maximum $\bullet$.

Where EM starts can make a big difference.
Search from Many Different Starts

To minimize the effects of local maxima, you should search multiple times from different starting points.

MEME uses this idea:

- Start at many points
- Run for one iteration
- Choose starting point that got the “highest” and continue
Gibbs Sampling: One at a time

- Randomly change starting position of motif only one $s_i$ from all $t$ sequences
- Discard k-mer from those defining PFM and find new start for $s_i$
  - Make new PFM with remaining $t-1$ sequences
  - Choose start according to $p_i$
- Moves with more caution
- Can Move in direction away from local maximum
- More Detail…
Gibbs Sampling : Algorithm

- Select Starting positions

- Randomly choose a sequence, i, set it aside

- Create PFM from k-mers in t-1

- Only change position in chosen seq i
  - For each position, calc \( p_j \), that starting position generated by profile, P

- Choose new start in t, according to \( p_{i,...,n-k+1} \)

- Repeat
1. Start with random motif locations and calculate a motif model
2. Randomly select a sequence, remove its motif and recalculate temporary model
3. With temporary model, calculate probability of motif at each position on sequence
4. Select new position based on this distribution
5. Update model and Iterate
Gibbs Sampling and Climbing

Because gibbs sampling does not always choose the best new location, it can move to another place not directly uphill.

In theory, Gibbs Sampling less likely to get stuck at a local maxima.
EM : Motif Discovery

- Don’t know starting positions of motif
- Don’t know PFM
- Randomly start

Form the Profile, P from s

Find Probability of Sequence
  - Model at starting point \( j=1 \ldots n-l+1 \)
  - Choose Best Start for all \( t \)
  - Update model with new starting points
  - Repeat

Elaborations...
ZOOPS

- Zero Or One Motifs per sequence
- Need another parameter to describe
- Situation is that one sequence might not even have a motif
- So don’t use that sequence to make the profile, PFM
E-step in the ZOOPS Model

- We need to consider another alternative: the $i$th sequence doesn’t contain the motif
- We add another parameter (and its relative)

\[ \lambda \]
- prior prob that any position in a sequence is the start of a motif

\[ \gamma = (n - k + 1)\lambda \]
- prior prob of a sequence containing a motif
E-step in the ZOOPS Model

\[
P(Z_{ij} = 1) = \frac{\Pr(S_i \mid Z_{ij} = 1, M) \lambda}{\Pr(S_i \mid Q_i = 0, M)(1 - \gamma) + \sum_{j=1}^{n-k+1} \Pr(S_i \mid Z_{ij} = 1, M) \lambda}
\]

- Here \( Q_i \) is a random variable that takes on 0 to indicate that the sequence doesn’t contain a motif occurrence.

\[
Q_i = \sum_{j=1}^{n-k+1} Z_{i,j}
\]
M-step in the ZOOPS Model (Count)

- update $p$ same as before
- update $\lambda, \gamma$ as follows

\[
\lambda^{(t+1)} = \frac{\gamma^{(t+1)}}{n-k+1} = \frac{1}{(n-k+1)^* t} \sum_{t} \sum_{i=1}^{m} Z_{i,j}^{(t)}
\]

- average of $Z_{i,j}^{(t)}$ across all sequences, positions
TCM

- Zero or more motifs in a sequence

- So now, not only could a sequence have no motif, but could also have many!

- Adds dimension of searching subsequences, W
E-step in the TCM Model

- the TCM model treats each length $W$ subsequence independently
- to determine the likelihood of such a subsequence:

$$Z_{ij} = \frac{\Pr(S_{i,j} \mid Z_{ij} = 1, M) \lambda}{\Pr(S_{i,j} \mid Z_{ij} = 0, B)(1 - \lambda) + \Pr(S_{i,j} \mid Z_{ij} = 1, M) \lambda}$$

- subsequence isn’t a motif
- subsequence is a motif

- M-step same as before
Motif Search

- **E-M : Motif Search**
  - Find Profile without starting s or profile, P
  - May converge to local maxima

- **Gibbs Sampling**
  - Move more conservatively, only change one $s_i$ at a time
  - Can move out of local maxima

- **MEME**
  - EM implementation with multiple starts

- **ZOOPS**
  - EM with additional option of no motif in sequence $t$

- **TCM**
  - EM with additional option of zero or more motifs in sequence $t$